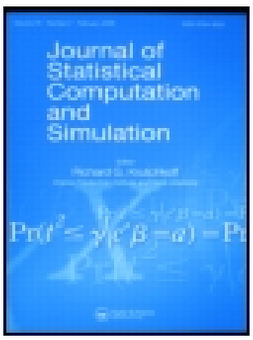

Journal of Statistical Computation and Simulation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

Comparison of methods of estimation for a goodness of fit test – an analytical and simulation study

Vimukthini Pinto & Roshini Sooriyarachchi

To cite this article: Vimukthini Pinto & Roshini Sooriyarachchi (2021): Comparison of methods of estimation for a goodness of fit test – an analytical and simulation study, Journal of Statistical Computation and Simulation

To link to this article: https://doi.org/10.1080/00949655.2021.1872078

Published online: 13 Jan 2021.

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION
https://doi.org/10.1080/00949655.2021.1872078

Comparison of methods of estimation for a goodness of fit test – an analytical and simulation study

Vimukthini Pinto and Roshini Sooriyarachchi

Department of Statistics, University of Colombo, Colombo, Sri Lanka

ABSTRACT

Multilevel modelling is a novel approach to analyse data which consist of a hierarchical or a nested structure. With advancements in multilevel modelling, there has been an advancement in the estimation techniques and also in goodness-of-fit tests which are vital to assess the fit of a model. However, these goodness-of-fit tests are not as yet tested to be suitable for models estimated using different estimation techniques. This study aims to conduct a comparison of methods of estimations for use in a goodness-of-fit test which is developed for binary response multilevel models. The comparison is based upon the mathematical background, extensive simulations and an application to a real-life dataset.

ARTICLE HISTORY
Received 18 August 2020
Accepted 2 January 2021

KEYWORDS
Estimation techniques; goodness-of-fit; marginal quasi likelihood (MQL); multilevel modelling; penalized quasi likelihood (PQL)

1. Introduction

1.1. Background of the study

1.1.1. Multilevel modelling

Frequently data arises with a hierarchy, a clustered or a nested structure attached with the data. This data arises from various fields such as, medical field where patients are nested within hospitals, educational field where students are nested within schools, family studies where children are nested within families etc. Similarly, multilevel data are the data structures which consists of two or more levels. Even though multilevel data can have more than two levels, similar to all the examples mentioned above, this study is only based upon the two-level multilevel structure and the study is carried out using the technique, multilevel modelling (MLM).

Multilevel models can be categorized based on the distribution of the response variable, type of data structure and the variance structure [1]. Considering the distribution of the response variable, the study considers the binary response variable with the logit model. The data structure used here is the simplest and the most common data structure, two-level hierarchical structure. Considering the variance structure, the study is based on the random intercept model where only the intercept is allowed to vary randomly. Thus, the model used throughout this study is the "random intercept, binary logistic multilevel

CONTACT Roshini Sooriyarachchi roshinis@hotmail.com; roshini@stat.cmb.ac.lk Department of Statistics, University of Colombo, Colombo 3, Sri Lanka

model" [2]. The more general random coefficient model is not used here for two reasons. Initially, we are interested in examining the effect of the Intra-Cluster Correlation (ICC) on the properties of the GOF test separately for each estimation method. For the variance components model (Random intercept model) the ICC also measures the proportion of the total variance which is between clusters. This intra-cluster correlation which is also the proportion of variance that is between clusters can easily be computed. In more complex models with random coefficients (The more general two level model) the intra unit correlation is not equivalent to the proportion of variance at the higher level [3]. Therefore, it is hard to determine. The second reason why a random coefficient model is not examined here is due to its many parameters and the complicated form of its covariance matrix [3]. Our results show that even for the simple random intercept model there are problems of convergence. This will be much more serious for the random coefficients model.

1.2. Estimation methods

There are several methods for estimation of parameters in multilevel models. Hox (4)[4] has mentioned several estimation procedures such as maximum likelihood method, generalized least squares method and generalized estimating equations. Also, there are Bayesian methods such as Markov Chain Monte Carlo (MCMC) [2].

However, there are more complex estimation procedures for logistic models. The most commonly used approach, quasi-likelihood approach, is to approximate the nonlinear link function by using a nearly linear link and include the effects of multilevel modelling. Following this approach, in this study, four estimation procedures considered are namely;

(1) Marginal Quasi Likelihood – Order 1 (MQL1)
(2) Marginal Quasi Likelihood – Order 2 (MQL2)
(3) Penalized Quasi Likelihood – Order 1 (PQL1)
(4) Penalized Quasi Likelihood – Order 2 (PQL2)

Researchers such as Rodriguez and Goldman (5) [5], Goldstein and Rasbash (6) [6] and Courgeau and Goldstein (7) [7] have found out the behaviour of these methods with varying multilevel structures. However, there is no evidence of one best estimation technique.

1.3. Goodness-of-fit (GOF) tests

Similar to model fitting, checking the GOF of the model is also an important step as statistical models are of no use if they provide the user with misleading results. If the models do not fit the data but are blindly used, the results will be erroneous and this might lead to biased conclusions. Moreover, it is equally important to make sure that the GOF test performs well and the test only recommends correct models. Therefore, it is necessary to use a test with a proven performance and a test which is only applicable to the model under consideration.

With advancements in multilevel modelling, there has been an evolution of GOF tests for varying multilevel structures. Chen (8) [8] describes some of the GOF tests available for two level Multilevel Binary Responses. Sturdivant (9) [9] and Sturdivant and

Hosmer (10) [10] have extended the Hosmer Lemeshow statistics for the single level logistic model by using smoothed residuals in the multilevel logistic model. Using the moments of these statistics they developed a standardized test statistic which should have an asymptotic standard normal distribution if the model was correctly specified. However, Sturdivant (9) [9] himself showed that this Normality assumption was problematic at the tails. Chen 8 [8] in his Ph.D. thesis has suggested to replace the Normal approximation by the scaled chi-square distribution with the same moments. Using some limited simulation studies he has shown that this approach is successful. However, probably due to some limitations in calculating the smoothed residuals, he has never published his work in a research journal. The GOF test proposed by Perera, Sooriyarachchi and Wickramasuriya (11) [11] for two level binary logistic models marks a milestone in GOF tests by acting as the base for more advanced GOF tests. However, the goodness of fit of the test was only examined by considering models estimated using PQL-2 method.

1.4. Objective of the study

Due to the above-mentioned limitations in the first two GOF testing methods described in section 1.3 these two methods were not considered for further study. As the method proposed by Perera, Sooriyarachchi and Wickramasuriya (11) [11] has no such limitations the current study was based only on this GOF test. Here, as there exists no best method of estimation, one can conduct estimations in multilevel models using any estimation method. Perera, Sooriyarachchi and Wickremasuriya [11] in the development of a GOF test for multilevel binary responses used only PQL 2 as the method of estimation in examining the properties of the developed GOF test. They did not examine the other methods of quasi likelihood available. However, the developed GOF test might be applicable/inapplicable for the other methods. In this research we examine all four methods of quasi likelihood estimation available in MLwiN and provide recommendations as to what method is most appropriate in each situation. Further, the simulation results are backed up by analytical proofs. This is the value added over Perera, Sooriyarachchi and Wickremasuriya [11]. Therefore, the aim of this study is to compare the four main methods of estimations for use in this GOF test and to recommend when and where this test can be used with respect to estimation methods.

2. Literature review

As multilevel data generates routinely in numerous fields, multilevel modelling has become a popular field of research during the last two decades. There exists evidence on multilevel analysis from the era of 1980s in different fields of research. Mason, William, George and Entwisle (1983) [12], Blalock (13)[13] Jackson (14)[14] and Goldstein (1995)[3] are few of the pioneers in multilevel data analysis. A clear introduction of the logic and statistical theory behind multilevel models can be credited to Steenbergen and Jones (15)[15] Understanding the practical importance, Peugh (16)[16] designed an article with the goal to explain the major decision-making steps necessary to enable applied researchers to conduct, interpret, and present the results of multilevel modelling.

Due to the wide usage of multilevel modelling, there has been an advancement in multilevel modelling concepts as well. This section is mainly designed to discuss the literature behind methods of estimations and the GOF tests for multilevel models.

2.1. Methods of estimations

Estimation of parameters is a vital step in statistical modelling procedure. Accordingly, parameter estimation in multilevel models is also an important step to which should be given a higher level of attention. There exists a long history on development of estimation techniques in multilevel models. Thus, this section reviews the previous developments in estimation procedures.

Parameters of generalized linear models are primarily obtained by maximum likelihood method [17]. Similarly, multilevel models are also generally estimated using maximum likelihood method. But, according to Hox (4) [4], combining multilevel structure with generalized linear models, leads to complex models and estimation procedures. To overcome the issues caused by computationally intensive procedures, an approach called the quasi-likelihood approach was introduced by Wedderburn (18)[18]. As explained by Rasbash et al. (1)[1] and Hox (4)[4], when using this approach, the general procedure is to approximate the non-linear link by using the Taylor series. After the linearization, the model could be treated as a continuous model and the general estimation procedures such as IGLS or RIGLS applied.

When considering the Taylor series linearization, it leads to two main methods of approximations. When the Taylor series expansion uses only current estimated values of the fixed part, it is referred to as marginal quasi-likelihood (MQL) as proposed by Goldstein (19)[19]. Breslow & Clayton (20) [20] introduced penalized (or predictive) quasi-likelihood (PQL) by improving it with the inclusion of the residual. As both these methods are developed by considering a Taylor series, order of the series should be specified. Most often only the 1st term or the 2nd term are considered in the linearization [3].

Over the past, several researchers such as Rodriguez and Goldman[5,21], Goldstein and Rasbash (6)[6] and Browne (22)[22] have conducted extensive simulations on the approximate methods for binary response models. Rodriguez and Goldman [5] discussed the results obtained by MQL method as a method which produces effects biased towards zero. Also, their study confirmed that MQL2 produces only a modest improvement over MQL1. Goldstein and Rasbash (6) with the intention of improving the results by Rodriguez and Goldman (5) introduced PQL2 method and illustrated that PQL shows a considerable improvement on the level-2 standard deviations. Another important finding from the study was, bias arises with the smaller number of level-1 units within the level-2 unit. This study also confirmed the adequacy of MQL method with smaller variances of the random component. Moreover, their study pointed out that there is a possibility of getting worse estimates from 2nd order methods in some circumstances. Therefore, one should be careful with the estimation method to be used with different multilevel structures. As a suggestion to improve estimates, Goldstein and Rasbash (6) [6] suggested to include subsequent terms of the Taylor series. Several improvements to both MQL and PQL procedures have been done over the past years. For a single source of extraneous variation for PQL, a correction factor for the estimates of variance component was introduced by Breslow and Lin

(23) [23]. They have also suggested a 1st order correction term for regression coefficients estimated by PQL method. Sutradhar and Rao (24)[24] introduced a four-moment-based MQL approach to provide consistent and also more efficient estimates for both the regression and over-dispersion parameters compared to the general MQL estimators. However, the trade-off between computational feasibility and improvement of estimates still requires exploration.

Goldstein and Rasbash (6) [6] suggested bootstrap methods or Gibbs Sampling to improve the quasi-likelihood estimates. Browne (22) [22] and Rodriguez and Goldman (21)[21] investigated approximate methods along with the bootstrap method and the Bayesian method. These simulations confirmed the improvement of estimates with bootstrap methods and Bayesian methods. However, computationally these proved to be difficult.

2.2. Assessing the model fit

Any fitted model should be evaluated and confirmed on its performance before making inferences from the developed model. This is basically to check if the predicted values from the model reflects the true outcome of the data. This is referred to as 'goodness of fit' (Hosmer, Lemeshow, & Sturdivant, 25)[25]. To assess the goodness of the fit, as per the explanation by Hosmer et al. (25)[25], there are three main approaches.

- Computation and evaluation of overall measures of fit
- Examination of the individual components on the summary statistics
- Examination of other measures of distance between observed and fitted values

This study only considers the 1st point, assessing the fit of the model through an overall measure of fit. Due to the presence of infinitely many numbers of points in continuous data, GOF tests developed for continuous models cannot be implemented for discrete data. Moreover, due to the hierarchy present in multilevel data, the GOF test for single level data cannot be implemented with multilevel models. Following is the theory behind the GOF test which is evaluated in this study for the test proposed by Perera et al. (11) [11] for binary logistic multilevel models.

2.3. Goodness of fit test by Perera et al. (11)

By taking the understanding from Hosmer-Lemeshow (1980) [26] test for single level binary data and the test proposed by Lipsitz Fitzmaurice and Molenberghs et al. (1996) [27] for single level ordinal data, Perera et al. (11) [11] made the advancement to the multilevel structure. The procedure of the test is as follows.

Considering k number of clusters and n_j number of observations within each cluster, the multilevel logistic regression model for binary data is fitted as the initial step. The two-level random intercept model for the probability of success (π_{ij}) using the logit link function with a single explanatory variable x_{ij} is given by,

$$\log i(\pi_{ij}) = \beta_{0j} + \beta_1 x_{ij} \tag{1}$$

where $\beta_{0j} = \beta_0 + u_{0j} \sim N(0, \sigma^2_{u0})$.

The parameters of this model are obtained by considering the PQL 2 method. The probability of success $\hat{\pi}_{ij}$ for i^{th} observation in the j^{th} cluster is estimated from the fitted model.

The next step is to apply the Hosmer-Lemeshow test. However, due to the nested structure of the data, it is not possible to ignore the clustering and apply the test as it is. Therefore, these estimated probabilities are sorted in ascending order and are collapsed in to G groups within each cluster. By taking the general partitioning of $G = 10$ groups recommended as in the Hosmer-Lemeshow test, the percentile-based grouping strategy is applied to each cluster.

After the partitioning of the data, the *(G-1)* indicator variables should be defined as in the test by Lipsitz et al. (27) [27]. Indicator variables are defined within each cluster such that;

$$I_{gij} = \begin{cases} 1; & \text{if } \hat{\pi}_{ij} \text{ is in region } g \\ 0; & \text{otherwise} \end{cases}$$

where $g = 2, 3, \ldots, G$.

The dataset is sorted back to its previous format and the alternative model is fitted in a similar manner as in the test by Lipsitz et al. (27) [27]. The alternative model is;

$$logit(\pi_{ij}) = \beta_{0j} + \beta_1 x_{ij} + \sum_{g=2}^{G} I_{gij}\gamma_g \tag{2}$$

$$where \quad \beta_{0j} = \beta_0 + u_{0j} \quad and \quad u_{0j} \sim N(0, \sigma^2_{u_0})$$

$$\sum_{g=2}^{G} \gamma_g I_{gij} = \gamma_2 I_{2ij} + \gamma_3 I_{3ij} + \ldots + \gamma_G I_{Gij}$$

$i = 1,2,\ldots,nj$ and $j = 1,2,\ldots,k$ where k is the number of clusters.

After fitting this alternative model, the joint Wald statistic is calculated to check the following hypothesis.

H_0: $\gamma_2 = \cdots = \gamma_G = 0$ vs

H_a: at least one coefficient of the indicator variables is not zero

Under the null hypothesis, the joint Wald statistic is assumed to follow a χ^2 distribution with *G-1* degrees of freedom. Thus, if the calculated statistic is greater than the 5% value of the χ_{G-1} distribution, it indicates a lack of fit of the model.

3. Methodology

The methods followed to analyse the applicability of the GOF test for the four methods of estimations are in three forms; the simulation study, analytical study and the application to the practical dataset.

3.1. Simulation study

To assess the performance of the GOF test with varying levels of cluster sizes, varying levels of observations within each cluster, varying values for ICC and varying methods of

estimation, simulations are carried out. To check the performance of the GOF test, the type I error should be checked and the power value should be determined. GOF test can be considered as applicable if the Type I error rate resides within the limits and if the test yields a considerably high-power value.

The two-level multilevel model is considered when developing simulations for both type I error and power. Under each scenario, 1000 datasets are generated using macros in MLwiN version 2.19. To ensure that the methods of estimations are done on the same dataset under the specified scenario, seed value of 100 is set on each simulation. The value 100 is selected based upon the seed value specified by Rasbash et al. (2017) [1]

Recalling the equation for binary logistic multilevel models, to simulate the explanatory variable, Perera et al. (11) [11] have suggested to use either Bernoulli distribution, normal distribution or uniform distribution. Hence, following Perera et al. (11) [11], the explanatory variable is simulated from a normal distribution with mean of 2.0 and standard deviation of 1.0.

Equations (1) and (2) are the two types of models fitted under the simulations study. These models are fitted under various conditions and the GOF of model (1) is assessed under those conditions by using the joint Wald statistic (Perera et al., 11) [11] with the following hypothesis of interest.

H_o: $\gamma_2 = \gamma_3 = \cdots = \gamma_{10} = 0$

H_a: at least one coefficient of the indicator variables is not zero

Then the joint Wald statistic obtained from MLwiN is compared with the chi- square value of 9 degrees of freedom at 5% significance level. The 9 degrees of freedom is selected as there are 10 groups leading to 9 independent quantities taken in to account.

3.1.1. Factors considered for simulations

According to the Equation (1) mentioned above, the model consists of a fixed slope parameter β_1, fixed component of the intercept β_0, and the random component of the intercept u_{0j} which should be determined. The usual practice is to set out the parameters after conducting a trial and error procedure. To maintain consistency with the simulations conducted by the developer of the test, Perera et al. (11) [11], and as these values are obtained from a trial and error method, parameter values are considered to be,

To assess Type I error: $\beta_0 = -0.686$, $\beta_1 = 0.707$

To assess power: $\beta_0 = -0.686$, $\beta_1 = 0.3535$

The sample size associated with the model depends on two main criteria. The number of clusters present in the study and the number of observations per each specified cluster. According to the guidelines specified by Maas and Hox (28) [28], Kreft and de Leeuw (29) [29] and Perera et al. (11) [11] four scenarios of sample sizes are selected and are given in Table 1, (Pinto & Sooriyarachchi, 2) [2].

Three combinations of standard deviation values are selected in accordance with the values set out by Perera et al. (11)[11]. As standard deviation is used to determine the intra cluster correlation (ICC) by the equation, $\rho(logit) = (\sigma^2_{between}/\sigma^2_{between} + \frac{\pi^2}{3})$ (Rasbash et al., 1 [1]), the combinations of ICC values considered for the simulations are given in Table 2.

Table 1. Combination of Sample Sizes.

Case	Cluster Size	No: of Clusters	Sample Size
1	20	15	300
2	50	15	750
3	20	60	1200
4	50	60	3000

Table 2. Combinations of ICC.

Case	Standard Deviation	Variance	ICC
1	**1**	1	0.2331
2	**1.5**	2.25	0.4061
3	**2**	4	0.5487

Table 3. Combinations Considered in the Study.

Factor	Number of combinations	Values
Standard Deviation of the Random Component	3	1, 1.5, 2
Number of Clusters	2	15, 60
Number of Observations in each Cluster (Cluster Size)	2	20, 50
Method of Approximation	2	MQL, PQL
Order of the Taylor Series	2	1, 2
Total Combinations	$3 \times 2 \times 2 \times 2 \times 2 =$ **48**	

These three combinations are also comparable with the desirable ranges for ICC specified by Knox and Chondros (2004)[30] where per their research conducted on ICC values of the survey conducted in Australia, the practical values of ICC ranges between 0.06 and 0.45. Therefore, scenario 1 and 2 above are selected to be within the practical range and scenario 3 is selected to analyse the behaviour of the test for the larger ICC situation.

After establishing all the above-mentioned conditions, the estimation procedure and the linearization is to be specified. Thus, the combinations resulting from the choices made above are summarized in the following Table 3. Datasets are simulated for all the listed combinations here to determine both type I error and power.

3.1.2. Study of the type I error

Type I error occurs as a result of rejecting the null hypothesis when it is actually true. Therefore, after generating data under the correct null hypothesis for 1000 datasets, the number of times it rejects the null hypothesis is obtained and it is checked whether it is within the 95% probability interval for α (Probability of making a type I error). To account for the random variation, probability levels for the Type I error is developed for 1000 datasets as (0.036, 0.064). Thus, the Type I error is calculated for all the 48 combinations listed previously and it is checked if the calculated value resides within this band. In cases of non-convergence of some of the 1000 datasets, new confidence intervals are calculated by considering only converging datasets. The simulation results for each method of estimation are given in Tables A1–A4 (Pinto & Sooriyarachchi, 2 [2]), in Appendix A.

The 48 simulations conducted for type I error is grouped in to four main sections with respect to the method of estimation. According to the results obtained, estimations done by MQL1 method for all the combinations produce type I errors outside the

acceptable range while PQL2 produces acceptable values for almost all the combinations. PQL1 produces considerably good results for largest sample size while MQL2 seems to produce better results when the ICC is low as all the Type I errors are within limits or extremely boarder line for ICC = 1. While the former result is well known the latter result has also been found by Guo and Zhao (2000) [31]. Run times associated with the four methods for almost all the combinations are produced according to a pattern where MQL1 < MQL2 < PQL1 < PQL2.

Considering the convergence issues, only five convergence problems are encountered out of 48 combinations and all five are observed when the number of clusters is small (n = 15). Order-1 methods seem to produce convergence issues when the ICC is at its lowest. Both these issues are present with the lowest sample size (k = 15, n = 20). When the ICC is moderate, two convergence issues are present with the MQL2 procedure when the number of clusters is small. For the extreme ICC, only one convergence issue is present with MQL2 with the smallest sample size. It should also be noted that the majority of convergence issues are present with the MQL method.

Following figure, Figure 1 provides a graphical representation of the Type I error results.

As the Figure 1 illustrates, Type I errors seems to inflate when the number of clusters is high (k = 60) for MQL2 estimation method. Also there exists a pattern when k = 60 for all the standard deviations under consideration, where type I errors for MQL1 < PQL1 < PQL2 < MQL2. Only type I errors produced by PQL2 method always resides within the limits while MQL1 is always below the limit. Considering the lowest standard deviation, there seems a similar pattern of type I errors for all the sample sizes, where type I errors for MQL1 < PQL1 < PQL2 < MQL2. However, no such pattern is visible with other standard deviations. In conclusion, GOF test seems to be suitable for almost all the combinations fitted using PQL2 method. However, for models fitted using MQL1, GOF test always produce Type I errors below the boundary.

3.1.3. Study of power

The power of a test is the probability of rejecting the false null hypothesis. Thus, this section focuses upon the simulations carried out in order to evaluate the power under each of the situations listed under Table 3. Unlike the simulations under type I error, two sets (1000 each) of datasets are simulated by considering the explanatory variable due to the reason provided below.

Set 1: 1000 datasets are simulated such that $x_{ij} \sim N(2,1)$

Set 2: 1000 datasets are simulated such that $x_{ij} \sim N(2,4)$

Figure 1. Type I Error Simulation Results.

The second set is generated to improve the power values which are generated from first set. As Perera et al. (11) [11] pointed out, when the random effect is larger than the covariate effect, the explanatory power of the explanatory variable is reduced. Thus, the 2nd set is simulated with a variance of 4.0. Similar to the simulations under type I error, all the datasets are simulated using macros in the MLwiN software. However, data is generated from an incorrect null hypothesis and the probability of rejection of the null hypothesis is obtained to calculate the power. The incorrect form of the null hypothesis used is,

$$\log i(\pi_{ij}) = \beta_{0j} + \beta_1 \ln (x_{ij})^2$$

where $\beta_{0j} = \beta_0 + u_{0j}, u_{0j} \sim N(0,\sigma^2_{u0}), i = 1,2,\ldots,nj$ and $j = 1,2,\ldots,k$ where k is the number of clusters

Then the fitted probabilities can be calculated as,

$$\pi^{ij} = exp(\beta_0 + u_{0j} + \beta_1 \ln (x_{ij})^2)/1 + \exp(\beta_0 + u_{0j} + \beta_1 \ln(x_{ij})^2)$$

Following tables, Tables B1–B4 provide the simulation results obtained (Pinto & Sooriyarachchi, 2 [2]). These are given in Appendix B.

As per the results obtained, there is a large increment in power value with the increase in standard deviation of the explanatory variable. Another interesting finding is regarding the convergence of the datasets. 90% of the datasets did not encounter any issues of convergence. However, when the cluster size is small and when the ICC is high, these issues are encountered with the MQL 2 procedure. Moreover, there is an interesting relationship with the runtime as well. PQL methods show a general tendency to take higher runtimes and PQL2 takes the highest time. Runtimes taken by non-converging combinations are considerably lower. Figure 2 gives a graphical representation for the power results obtained when the explanatory variable is simulated with variance of 4.0.

It could be seen that; higher power is obtained with the larger number of clusters which is in accordance with the reasoning done by Schoeneberger (31) [31]. For an instance, combination 3 (k = 60, n = 20) has 1200 observations while combination 2 (k = 15, n = 50) has only 750 observations. Yet the power associated with combination 3 is only marginally higher than that associated with combination 2. The study of Schoeneberger (31)[31] also revealed on the considerable increase in mean power estimates as the level-2 sample size increased, with smaller differences noted across level-1 sample size.

Moreover, there seems a decrement in power with the increase in ICC for all the methods of estimations. In conclusion, MQL2 and PQL2 methods produce satisfactory power when

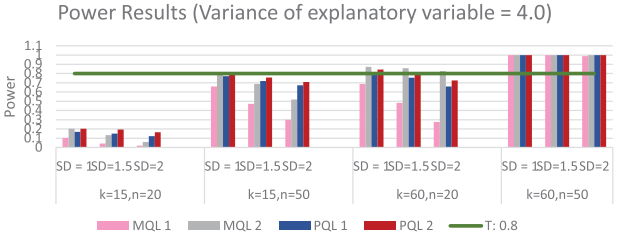

Figure 2. Power Simulation Results.

number of clusters is high. Additionally, MQL1 generates the lowest power irrespective of any other factor.

3.2. Analytical study

The mathematical background behind the impact of methods of estimations on the properties of GOF tests is a new area of research. Thus, this section of the study is developed to do a theoretical analysis of the impact of methods of estimations on the goodness of fit test for binary multilevel models.

Equation (3) gives the definition of the Taylor series expansion. (3)

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f^{(3)}(a)}{3!}(x-a)^3 + \ldots + \frac{f^{(n)}(a)}{n!}(x-a)^n + \ldots$$

By using Taylor series expansion, $(t+1)^{th}$ iteration of the iterative generalized least squares (IGLS) algorithm is given by (Considering only up to second order),

$$f(K_{t+1}) = f(K_t) + X_{ij}(\hat{\beta}_{t+1} - \hat{\beta}_t)f'(K_t) + u_j f'(K_t) + u_j^2 f''(K_t)/2 \tag{4}$$

where

$$f'(K_t) = f(K_t)[1 + \exp(K_t)]^{-1} \tag{5}$$

$$f''(K_t) = f'(K_t)[1 - \exp(K_t)]\,[1 + \exp(K_t)]^{-1} \tag{6}$$

Note that the 2nd term of (4) updates the fixed part of the model while 3rd term adjusts for the random component. Now, Equation (4) is a linear model and the procedures used for linear multilevel models can now be implemented.

Choices made on K_t defines the two main methods of approximation considered in this study, MQL (Equation a) and PQL (Equation b).

(a) $K_t = \boldsymbol{X}_{ij}\hat{\boldsymbol{\beta}}_t$
- uses only the fixed part of the Taylor expansion (MQL)

(b) $K_t = \boldsymbol{X}_{ij}\hat{\boldsymbol{\beta}}_t + \hat{u}_{t,j}$
- uses the fixed part of the Taylor expansion and current estimated residuals (PQL)

3.2.1. MQL method

MQL, the approach proposed by Goldstein (19) [18] is the method which is adopted when K_t uses only the fixed part of the Taylor series expansion. This method is generally considered to be appropriate when the interest is focused mainly on the marginal relationship between the response and the variables [32]. After the choice of MQL as the approximation method, the next choice that should be made is the order of linearization. The usual approach is to consider either the order-1 or order-2. The two Equations (7) and (8) are shown below to provide a clearer representation of MQL 1 and MQL 2 methods.

MQL 1:

$$f(K_{t+1}) = f(K_t) + X_{ij}\left(\hat{\beta}_{t+1} - \hat{\beta}_t\right)f'(K_t) + u_j f'(K_t) \tag{7}$$

MQL 2:

$$f(K_{t+1}) = f(K_t) + X_{ij}(\hat{\beta}_{t+1} - \hat{\beta}_t)f'(K_t) + u_j f'(K_t) + u_j^2 f''(K_t)/2 \tag{8}$$

In both the instances above, $K_t = \boldsymbol{X}_{ij}\hat{\boldsymbol{\beta}}_t = \hat{\beta}_{0t} + \hat{\beta}_{1t}x_{ij}$.

3.2.2. PQL method

PQL technique, as a remedy to the effect caused by random components of the model is proposed by Breslow and Clayton (20) [19]. In this procedure, expansion around $\hat{u}_{t,j}$ is also considered. Equations (9) and (10) explain this.

PQL1:

$$f(K_{t+1}) = f(K_t) + X_{ij}(\hat{\beta}_{t+1} - \hat{\beta}_t)f'(K_t) + (u_j - \hat{u}_j)f'(K_t) \tag{9}$$

PQL2:

$$f(K_{t+1}) = f(K_t) + X_{ij}(\hat{\beta}_{t+1} - \hat{\beta}_t)f'(K_t) + (u_j - \hat{u}_j)f'(K_t) + (u_j - \hat{u}_j)^2 f''(K_t)/2 \tag{10}$$

For both these instances, $K_t = \boldsymbol{X}_{ij}\hat{\boldsymbol{\beta}}_t + \hat{u}_{t,j} = \hat{\beta}_{0t} + \hat{u}_{t,0j} + \hat{\beta}_{0t}x_{ij}$

After obtaining the differences between the estimation methods, to assess the impact of these methods a derivation for type I error and power is obtained as follows.

Type I error:

$$H_o : \gamma_2 = \gamma_3 = \ldots = \gamma_{10} = 0$$

H_a : At least one coefficient of the indicator variables is not equal to zero

By considering the above hypothesis,

$$\begin{aligned}
\alpha &= \Pr(\textit{Rejecting } H_0 | H_0 \textit{ is true}) \\
&= \Pr(\textit{Multilevel binary logistic model does not fit the data well} | \textit{Model fits the data}) \\
&= \Pr(\textit{Joint Wald Statistic} > \chi^2_{9,5\%} | \gamma_2 = \gamma_3 = \ldots = \gamma_{10} = 0) \\
&= \Pr\left(\sum_{g=2}^{10}\left(\frac{\widehat{\gamma_g} - \gamma_g}{SE(\hat{\gamma}_g)}\right)^2 > \chi^2_{9,5\%} | \gamma_2 = \gamma_3 = \ldots = \gamma_{10} = 0\right)
\end{aligned}$$

For all the values of g, as all γ_g values are given to be zero,

$$\alpha = \Pr\left(\sum_{g=2}^{10}\left(\frac{\widehat{\gamma_g}}{SE(\hat{\gamma}_g)}\right)^2 > \chi^2_{9,5\%}\right) \tag{11}$$

At the 5% level of significance, as the null hypothesis check for 9 coefficient values, the critical value considered is $\chi^2_{9,5\%} = 16.919$. How α changes with the method of estimation can be argued as follows.

It is assumed that, $\hat{\gamma} \sim N(\gamma, \sigma_{\gamma}^2)$

As the estimates get better; $[E(\hat{\gamma}) - \gamma] \to 0$

When data is generated from the true null hypothesis, $E(\hat{\gamma}) \to \gamma \to 0$ ($\because \gamma = 0$ *under* H_0)

Thus, $\hat{\gamma} \to 0 \because E(\hat{\gamma}) \to 0$. Thus, by using this formulation, impact of methods of estimations on Equation 11 could be qualitatively compared.

Power:

$$\begin{aligned} Power &= 1 - \beta = 1 - \Pr(\textit{Type II error}) = 1 - \Pr(\text{ Not rejecting } H_0 \mid H_0 \textit{ is false}) \\ &= \Pr(\textit{Rejecting } H_0 \mid H_0 \textit{ is false}) \\ &= \Pr(\textit{Multilevel model does not fit the data well} \mid \textit{Model does not fit the data}) \\ &= \Pr(\textit{Joint Wald Statistic} > \chi^2_{9,5\%} \mid \textit{at least one coefficient of the indicator} \\ &\qquad \textit{variables is not zero}) \end{aligned}$$

$$Power = \Pr\left(\sum_{g=2}^{10}\left(\frac{\widehat{\gamma_g} - \gamma_g}{SE(\hat{\gamma}_g)}\right)^2 > \chi^2_{9,5\%} \mid \gamma_2 \neq 0 \cup \gamma_3 \neq 0 \ldots \cup \gamma_{10} \neq 0\right) \quad (12)$$

Equations (7)–(12) implies that the parameter estimate for a specific explanatory variable for MQL2 and thus the fitted value, the indicator variable and the significance of the indicator variables will usually be larger than that for MQL1. Similarly, the parameter estimates for a specific explanatory variable for PQL2 and thus the fitted value, the indicator variable and the significance of the indicator variables will usually be larger than that for PQL1. The $\widehat{\gamma_g}$'s are the coefficients associated with the indicator variables. Under the alternative hypothesis at least one of the $\gamma_g{}'$s is non-zero. Thus, the power associated with MQL2 will usually be larger than the power associated with MQL1 and the power associated with PQL2 will usually be larger than the power associated with PQL1 for the respective combinations. This is further highlighted in the previous simulations.

3.3. Practical application

To compare the practical results obtained by the GOF test with the change in estimation procedure, and also to compare the parameter estimates given by each method of estimation, models are fitted using the four methods of estimations. Next the GOF test is applied to the models to compare the recommendations given by the test under each method.

The inbuilt dataset used, a sub sample of '1989 Bangladesh fertility survey dataset' [33] consists of 2687 records of data collected from Bangladesh women over the country nested within their district of residence. The dataset comprises of a binary response variable, 'use' which indicates whether the individual uses contraceptives at the time of data collection or not. As discussed in the previous section, dataset comprises of 2687 women nested within 60 districts. Districts are coded from 1 to 61 where no observations are reported from district 54. Thus, a total of 60 districts are available in the dataset. However, number of women belonging to some districts are less than 10. Thus, to avoid any complications in applying the GOF test with the use of 10 indicator variables, these districts are ignored in the model fitting. Moreover, districts which contain women between 10 and 20 are also ignored to avoid any possibilities of non-convergence.

Table 4. Summary of Four Methods of Estimations.

	MQL 1			MQL 2			PQL 1			PQL 2		
	Est	SE	Z-Ratio	Est	SE	Z-Ratio	Est	SE	Z-Ratio	Est	SE	Z-Ratio
Cons	−1.553	0.250	−6.2120	−1.616	0.254	−6.3622	−1.604	0.256	−6.2656	−1.6260	0.259	−6.2780
lc1	1.145	0.135	8.4815	1.189	0.137	8.6788	1.181	0.137	8.6204	1.197	0.138	8.6739
lc2	1.437	0.148	9.7095	1.493	0.150	9.9533	1.483	0.151	9.8212	1.504	0.151	9.9602
lc3	1.479	0.154	9.6039	1.537	0.156	9.8526	1.525	0.157	9.7134	1.546	0.157	9.8472
educ2	0.229	0.131	1.7481	0.239	0.132	1.8106	0.236	0.132	1.7879	0.239	0.132	1.8106
educ3	0.665	0.144	4.6181	0.693	0.145	4.7793	0.687	0.146	4.7055	0.697	0.147	4.7415
educ4	1.142	0.127	8.9921	1.188	0.129	9.2093	1.183	0.13	9.1	1.199	0.131	9.1527
urban	0.516	0.105	4.9143	0.537	0.106	5.0660	0.532	0.107	4.972	0.539	0.107	5.0374
hindu	0.405	0.128	3.1641	0.423	0.129	3.2791	0.417	0.13	3.2077	0.423	0.131	3.2290
age	−0.018	0.007	−2.5714	−0.019	0.007	−2.7143	−0.018	0.007	−2.5714	−0.019	0.007	−2.7142
d.pray	−0.999	0.499	−2.0020	−1.038	0.504	−2.0595	−1.02	0.512	−1.9921	−1.039	0.519	−2.0019
σ_u^2	0.185	0.057	3.2456	0.189	0.058	3.2586	0.196	0.06	3.2667	0.203	0.062	3.27414

Next, binary logistic multilevel models are fitted under each method of estimation. In order to identify the most important variables, forward selection is implemented with the use of Wald statistic at 5% level of significance. Table 4 summarizes the coefficients and z-ratios for the best models under each estimation method.

Next, the GOF test is applied for the above-mentioned models and interestingly the test indicated that the models fit the data well by providing chi- square values more than the critical value. To assess the suitability of the test, 1000 datasets were simulated to closely match the real dataset and this is explained in section 3.4.

3.4. Simulation to match the Dataset

A simulation is conducted to approximate the type I error and power of the test for each method of estimation before making recommendations from the test. To closely match the dataset which comprises of 2711 units with 49 districts, a balanced dataset is simulated with a sample size of 2700 consisting of 45 clusters. Given in Table 5 are the parameters used in the simulation.

The values for parameters are taken to closely represent the four models to be fitted and distribution of each variable is based upon the model diagnostics obtained for each variable.

1000 datasets are generated from each method of estimation by using the correct form of the model to assess type I error. To assess power, another 1000 datasets are generated using an incorrect functional form as highlighted in section 3.1.3 using two explanatory variables under each method of estimation. Presented in Table 6 are the results obtained from the simulation.

Interestingly, results obtained from all four methods produce type I errors within the boundary (0.036, 0.064) and satisfactory powers from the test. However, the estimated Type I error for MQL1 is more towards the lower margin (just within the limits). To summarize the results from the practical example and the simulation based on the practical example, Table 12 shows that for the example the most significant random variation is given by PQL1 and PQL2 and approximately the z-statistic is 3.27. From tables A3 and A4 it can be seen that PQL1 and PQL2 give an estimate of type I error closest to 0.05 and high estimated power. Therefore, it is recommended that PQL1 and PQL2 are the most suitable methods

Table 5. Parameters to Match the Real-World Dataset.

Parameter	Value	Remarks
Number of clusters	45	To match the original cluster size 49
Cluster size	60	To match the sample size 2711 (60 is chosen for easy application of the test)
Standard error of the random component	0.4385	Mean standard error of four models (6.1), (6.2), (6.3) and (6.4)
Constant term	−1.6	Mean constant term of the four models
lc: [$Pr(lc1 = 1) = 0.1808, Pr(lc2 = 1) = 0.1612, Pr(lc3 = 1) = 0.3880$]		
- lc1	1.178	Mean lc1 of the four models
- lc2	1.479	Mean lc2 of the four models
- lc3	1.522	Mean lc3 of the four models
educ: [$Pr(educ2 = 1) = 0.0136, Pr(educ3 = 1) = 0.0948, Pr(educ4 = 1) = 0.1560$]		
- educ2	0.236	Mean educ2 of the four models
- educ3	0.686	Mean educ3 of the four models
- educ4	1.178	Mean educ4 of the four models
urban	0.531	Mean value of the four models [$Pr(urban = 1) = 0.2840$]
hindu	0.417	Mean value of the four models [$Pr(hindu = 1) = 0.1354$]
age	−0.019	Mean age of the four models [$Age \sim N(-0.411, 8.953)$]
d_pray	−1.024	Mean of the four models [$d\_pray \sim N(0.427, 0.023)$]

Table 6. Simulation for the Real Dataset.

Estimation Technique	Type I Error	Power
MQL1	0.036	0.98
MQL2	0.057	0.98
PQL1	0.050	0.98
PQL2	0.052	0.98

for analysing the example. Based on this example it could be recommended to practitioners that the Penalized Quasi Likelihood (PQL) methods are more appropriate for model fitting than Marginal Quasi Likelihood methods. Though PQL2 is usually the best method sometimes due to its complexity it might result in non-convergence. If so the simpler PQL1 can be used.

4. Discussion and conclusion

The aim of the study was to find the usage of the GOF test with varying methods of estimations. As a secondary objective, these four estimation methods were compared. Generally, PQL2 is considered as the procedure which produce the most unbiased estimates while MQL1 is the most biased. However, the simulations conducted by previous researchers such as Goldstein (3) [3] has indicated that the standard errors of the PQL2 estimates are comparatively higher, whereas MQL1 is comparatively lower. Results obtained for the practical dataset also supported that argument by producing higher standard errors for PQL methods while the highest was obtained for PQL2. Thus, as one best method of estimation is not available, depending on the structure of dataset, one can conduct the estimations in discrete response multilevel models by using any estimation method. Another issue in deciding a method of estimation, especially in simulation, is the presence of convergence problems in some methods of estimation.

The main conclusion from the study is that the GOF test developed by Perera et al. (11) [11] perform differently with the models estimated using the four methods of estimation.

The Type I error simulations and analytical study indicates that the test produces adequate Type I errors for models estimated using PQL2. However, the test fails to maintain the Type I error for models estimated using MQL1. For models estimated using MQL2 or PQL2, the test seems to produce adequate Type I errors depending on the sample size.

Considering the power of the test, it depends on the selected incorrect functional form. The incorrect form used was $ln(X^2)$. Thus, considering the form used in this study, the power of the test increase with the increase in sample size irrespective of the method of estimation and there seems an inverse relationship between power of the test and ICC. Moreover, the test produces better power values for models estimated using order-2 methods and by supporting the reasoning done by Schoeneberger (31) [31], power estimates seemed to increase as the level-2 sample size increase for all the methods of estimations.

The practical application indicates the use of the GOF test practically with any method of estimation. The performance of the test may however vary with the selected estimation method. The results of this study can be generalized up to a point in the sense that usually we can order the appropriateness in ascending order MQL1, MQL2, PQL1, PQL2. However, with small datasets there could be convergence problems with some methods and in this case some alternative method may be more suitable.

The study compares only the models fitted using the four main methods of estimation. However, there are more advanced methods such as bootstrap methods and MCMC methods. Moreover, the simulations are based upon balanced clusters, this could be extended for unbalanced clusters in the further research. An equation which comprises of all the properties such as number of clusters, cluster size, ICC and estimation method was not developed due to the complex iterative procedures. Therefore, numerical quantification of the effect caused by each estimation technique on Type I error and power was not looked at. Thus, it is beneficial to look in to more mathematical details. In our work with GOF testing for proportional odds multilevel models [34] it was seen that PQL2 gave convergence problems and thus PQL1 had to be used. In our experience with survival data [35] which is more complicated than discrete responses such as binary and ordinal categorical there were convergence problems with all the better estimation methods PQL2, PQL1 and MQL2 therefore the most basic method of MQL1 had to be used. Therefore, the conclusions from this study may be extended to other types of multilevel models though, no detailed study such as in this research has been done on other types of responses.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] Rasbash J, Steele F, Browne W, et al. A user's guide to MLwiN, version 3.00. Bristol, UK: Centre for Multilevel Modelling, University of Bristol; 2017.
[2] Pinto IV, Sooriyarachchi MR. Comparison of methods of estimation for use in goodness of fit tests for binary multilevel models. International Science Index 148, International Journal of Mathematical, Computational, Physical, Electrical and Computer Engineering. 2019;13(4):68–73.
[3] Goldstein H. Multilevel statistical models. 4th ed. UK: John Wiley and Sons, Ltd; 2011.
[4] Hox JJ. Multilevel analysis- Techiniques and applications. New York: Routledge; 2010.

[5] Rodriguez G, Goldman N. An assessment of estimation procedures for multilevel models with binary responses. Journal of the Royal Statistical Society. 1995;158(1):73–89.
[6] Goldstein H, Rasbash J. Improved approximations for mulitlevel models with binary respones. Journal of the Royal Statistical Society. 1996;159(3):505–513.
[7] Courgeau D, Goldstein H. (1997). Multilevel Statistical Models. *Population (French Edition)*.
[8] Chen, N.W. (2011). Goodness-of-fit test issues in generalized linear mixed models. Unpublished Ph.D. thesis, Texas A&M University.
[9] Sturdivant, R. X. (2005), "Goodness-of-Fit in hierarchical logistic regression models," Ph.D. Dissertation, University of Massachusetts Amherst, University Microfilms International, Ann Arbor.
[10] Sturdivant RX, Hosmer DW. A smoothed residual based goodness-of-Fit statistic for logistic hierarchical regression models. Comput Stat Data Anal. 2007;51:3898–3912.
[11] Perera A. P. , Sooriyarachchi M. R., Wickramasuriya, **S**. A goodness of fit test for the multilevel logistic model. Commun Stat Simul Comput. 2016;45(2):643–659.
[12] Manson WM, George YW, Entwisle B. (1983). Contextual Analysis Through the Multilevel Linear Model. In *Sociological Methodoly* (pp. 72–103).
[13] Blalock H. Contextual-effects models: theoritical and methodological isuues. Annu. Rev. Social. 1984;10:353–375.
[14] Jackson JE. Estimation of models with variable coefficients. Polit Anal. 1992;3:27–49.
[15] Steenbergen M, Jones B. Modeling multilevel data structures. Am J Pol Sci. 2002;46(1):218–237.
[16] Peugh J. A practical guide to multilevel modeling. J Sch Psychol. 2010., 48(1): 85–112.
[17] McCullagh P, Nelder JA. Generalized linear models. 2nd ed. London: Chapman and Hall; 1989.
[18] Wedderburn R. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika. 1974;61(3):439–447.
[19] Goldstein H. Nonlinear multilevel models, with an application to discrete response data. Biometrika. 1991;78(1):45–51.
[20] Breslow N, Clayton D. Approximate Inference in generalized linear Mixed models. J Am Stat Assoc. 1993;88(421):9–25.
[21] Rodriguez G, Goldman N. Improved estimation procedures for multilevel models with binary response: A case-study. Journal of the Royal Statistical Society, Series A. 2001;158(2):73–89.
[22] Browne WJ. Applying MCMC methods to multilevel models. Bath: University of Bath; 1998.
[23] Breslow NE, Lin X. Bias correction in generalized linear models with a single component of dispersion. Biometrika. 1995;82:81–92.
[24] Sutradhar BC, Rao RP. On marginal quasi-likelihood Inference in generalized linear Mixed models. J Multivar Anal. 2001;76:1–34.
[25] Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. 3rd ed. John Wiley & Sons, Inc; 2013, USA.
[26] Hosmer D, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. Commun Stat - Theory Methods. 1980;9(10):1043–1049.
[27] Lipsitz S, Fitzmaurice G, Molenberghs G. Goodness-of-fit tests for ordinal response regression models. Journal of the Royal Statistical Society. Series C (Applied Statistics). 1996;45(2):175–190.
[28] Maas C, Hox J. (2005). Sufficient sample sizes for multilevel modeling.
[29] Kreft IG, de Leeuw J. Introducing multilevel modeling. Newbury Park (CA): Sage; 1998.
[30] Knox and Chondros (2004).
[31] Schoeneberger JA. The impact of sample size and other factors when estimating multilevel logistic models. The Journal of Experimental Education. 2016;84(2):373–397.
[32] Liang K-Y, Zeger SL, Qaqish B. Multivariate regression Analyses for categorical data. Journal of the Royal Statistical Society. Series B (Methodological). 1992;54(1):3–40. http://www.jstor.org/stable/2345947.
[33] Huq NM, Cleland J. (1990). Bangladesh fertility survery, 1989. *Dhaka:National Institute of Population Research and Training (NIPORT)*.
[34] Epasinghe N, Sooriyarachchi R. A goodness of fit test for the multilevel proportional odds model. Commun Stat - Simul Comput. 2017;46(7):5610–5626.

[35] Balakrishnan K, Sooriyarachchi MR. A goodness of fit test for multilevel survival data. Commun Stat - Simul Comput. 2018;47(1):30–47.

Appendices

Appendix A – Results of Estimated Type I Error for each Method of Estimation

Table A1. Type-I error results for MQL 1

ICC	No: of Clusters (k)	Cluster Size (n)	Runtime	No: of Significant Datasets	Rejection Proportion	Results
Standard Deviation = 1	15	20	1m40s	5 from 387	0.013	Outside the limit
	15	50	1m50s	14	0.014	Outside the limit
	60	20	1m54s	11	0.011	Outside the limit
	60	50	2m12s	13	0.013	Outside the limit
Standard Deviation = 1.5	15	20	1m15s	7	0.007	Outside the limit
	15	50	1m14s	8	0.008	Outside the limit
	60	20	1m48s	0	0	Outside the limit
	60	50	2m34s	5	0.005	Outside the limit
Standard Deviation = 2	15	20	1m17s	4	0.004	Outside the limit
	15	50	1m17s	2	0.002	Outside the limit
	60	20	1m37s	2	0.002	Outside the limit
	60	50	2m02s	2	0.002	Outside the limit

Table A2. Type-I error results for MQL 2

ICC	No: of Clusters (k)	Cluster Size (n)	Runtime	No: of Significant Datasets	Rejection Proportion	Results
Standard Deviation = 1	15	20	2 m 10s	51	0.051	Within the limit
	15	50	1m52s	42	0.042	Within the limit
	60	20	2m28s	65	0.065	Outside the limit
	60	50	3m12s	65	0.065	Outside the limit
Standard Deviation = 1.5	15	20	2m11s	7 from 414	0.017	Outside the limit
	15	50	1m17s	10 from 238	0.042	Within the limit
	60	20	2m38s	65	0.065	Outside the limit
	60	50	3m05s	86	0.086	Outside the limit
Standard Deviation = 2	15	20	1m28s	5 from 422	0.012	Outside the limit
	15	50	2m13s	18	0.018	Outside the limit
	60	20	3m05s	82	0.082	Outside the limit
	60	50	4m13s	132	0.132	Outside the limit

Table A3. Type-I error results for PQL 1

ICC	No: of Clusters (k)	Cluster Size (n)	Runtime	No: of Significant Datasets	Rejection Proportion	Results
Standard Deviation = 1	15	20	2m02s	12 from 387	0.031	Outside the limit
	15	50	2m10s	30	0.03	Outside the limit
	60	20	2m49s	29	0.029	Outside the limit
	60	50	2m23s	38	0.038	Within the limit
Standard Deviation = 1.5	15	20	2m03s	32	0.032	Outside the limit
	15	50	2m16s	23	0.023	Outside the limit
	60	20	3m05s	29	0.029	Outside the limit
	60	50	2m50s	36	0.036	Within the limit
Standard Deviation = 2	15	20	2m28s	32	0.032	Outside the limit
	15	50	2m28s	38	0.038	Within the limit
	60	20	3m17s	24	0.024	Outside the limit
	60	50	4m09s	36	0.036	Just within the limit

Table A4. Type-I error results for PQL 2

ICC	No: of Clusters (k)	Cluster Size (n)	Runtime	No: of Significant Datasets	Rejection Proportion	Results
Standard Deviation = 1	15	20	2m40s	51	0.051	Within the limit
	15	50	2m55s	42	0.042	Within the limit
	60	20	3m33s	47	0.047	Within the limit
	60	50	4m42s	45	0.045	Within the limit
Standard Deviation = 1.5	15	20	2m05s	39	0.039	Within the limit
	15	50	2m58s	38	0.038	Within the limit
	60	20	4m23s	42	0.042	Within the limit
	60	50	3m20s	42	0.042	Within the limit
Standard Deviation = 2	15	20	3m41s	41	0.041	Within the limit
	15	50	3m55s	47	0.047	Within the limit
	60	20	4m38s	33	0.033	Outside the limit
	60	50	5m44s	39	0.039	Within the limit

Appendix B – Results of Estimated Power for each Method of Estimation

Table B1. Power results for MQL 1

			$x_{ij} \sim N(2,1)$		$x_{ij} \sim N(2,4)$	
ICC	No: of Clusters (k)	Cluster Size (n)	Runtime	Rejection Proportion	Runtime	Rejection Proportion
Standard Deviation = 1	15	20	2m12s	0.027	1m40s	0.103
	15	50	1m22s	0.06	1m17s	0.66
	60	20	1m31s	0.086	1m42s	0.688
	60	50	1m55s	0.445	2m11s	1
Standard Deviation = 1.5	15	20	1m07s	0.009	1m12s	0.043
	15	50	1m11s	0.025	1m19s	0.473
	60	20	1m33s	0.043	1m39s	0.482
	60	50	2m13s	0.266	2m04s	0.999
Standard Deviation = 2	15	20	1m19s	0.002	1m31s	0.02
	15	50	1m20s	0.011	1m27s	0.299
	60	20	1m37s	0.012	1m43s	0.277
	60	50	1m53s	0.121	2m15s	0.989

Table B2. Power results for MQL 2

			$x_{ij} \sim N(2,1)$		$x_{ij} \sim N(2,4)$	
ICC	No: of Clusters (k)	Cluster Size (n)	Runtime	Rejection Proportion	Runtime	Rejection Proportion
Standard Deviation = 1	15	20	1m48s	0.071	1m33s	0.202
	15	50	5m12s	0.183	1m44s	0.803
	60	20	2m05s	0.26	2m34s	0.873
	60	50	2m33s	0.683	3m05s	1
Standard Deviation = 1.5	15	20	2m540s	0.057	1m10s	0.132
	15	50	2m42s	0.132	2m01s	0.688
	60	20	2m41s	0.292	2m34s	0.857
	60	50	3m44s	0.681	3m19s	1
Standard Deviation = 2	15	20	40s	0.016	1m00s	0.059
	15	50	2m18s	0.084	2m29s	0.518
	60	20	2m53s	0.303	3m05s	0.827
	60	50	3m22s	0.633	3m40s	1

Table B3. Power results for PQL 1

			$x_{ij} \sim N(2,1)$		$x_{ij} \sim N(2,4)$	
ICC	No: of Clusters (k)	Cluster Size (n)	Runtime	Rejection Proportion	Runtime	Rejection Proportion
Standard Deviation = 1	15	20	1m47s	0.051	2m24s	0.168
	15	50	1m56s	0.127	2m29s	0.771
	60	20	2m19s	0.153	2m48s	0.788
	60	50	2m58s	0.609	3m29s	1
Standard Deviation = 1.5	15	20	3m07s	0.051	2m17s	0.15
	15	50	3m12s	0.12	2m27s	0.718
	60	20	3m10s	0.135	3m11s	0.755
	60	50	4m04s	0.558	3m48s	1
Standard Deviation = 2	15	20	2m38s	0.046	3m11s	0.122
	15	50	2m44s	0.116	3m03s	0.672
	60	20	3m08s	0.133	3m53s	0.659
	60	50	4m02s	0.514	4m49s	1

Table B4. Power results for PQL 2

			$x_{ij} \sim N(2,1)$		$x_{ij} \sim N(2,4)$	
ICC	No: of Clusters (k)	Cluster Size (n)	Runtime	Rejection Proportion	Runtime	Rejection Proportion
Standard Deviation = 1	15	20	2m26s	0.071	2m48s	0.203
	15	50	2m05s	0.161	3m17s	0.803
	60	20	3m11s	0.203	3m41s	0.844
	60	50	4m04s	0.631	4m45s	1
Standard Deviation = 1.5	15	20	2m53s	0.066	3m03s	0.194
	15	50	2m57s	0.158	3m11s	0.756
	60	20	4m36s	0.179	5m03s	0.794
	60	50	6m02s	0.589	6m12s	1
Standard Deviation = 2	15	20	3m52s	0.063	4m23s	0.164
	15	50	4m00s	0.153	4m33s	0.709
	60	20	4m22s	0.173	5m30s	0.726
	60	50	5m23s	0.54	7m42s	1